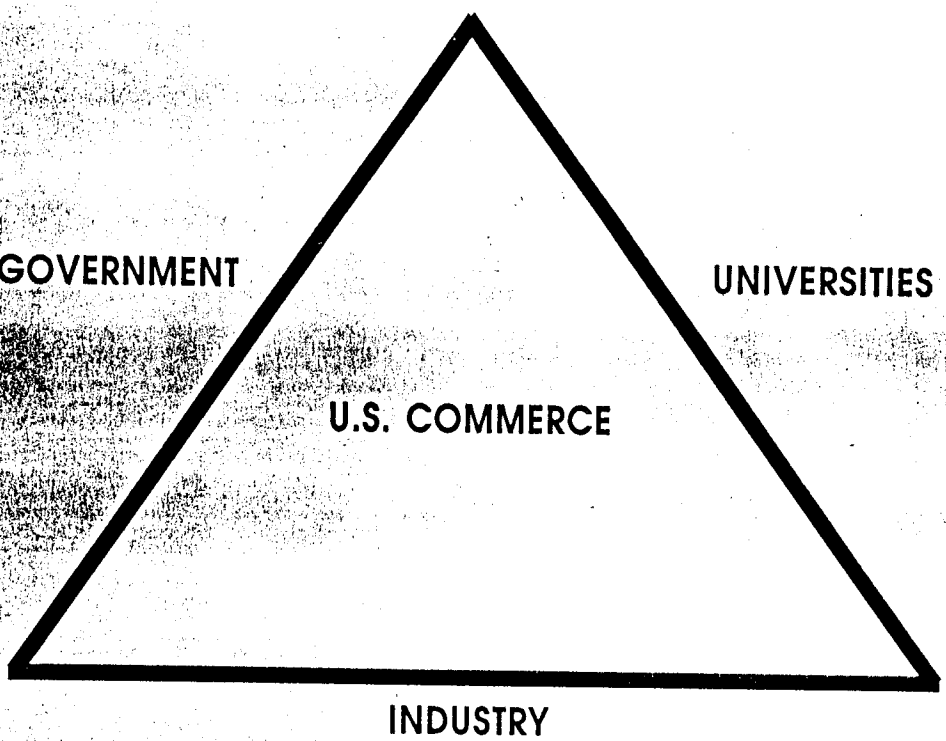
94-12

AIP CONFERENCE PROCEEDINGS 325

CONFERENCE ON NASA CENTERS FOR COMMERCIAL DEVELOPMENT OF SPACE

ALBUQUERQUE, NM 1995



EDITORS: MOHAMED S. EL-GENK
RAYMOND P. WHITTEN

DYNAMIC ROUTING IN HYBRID NETWORKS WITH INTEGRATED VOICE AND DATA TRAFFIC

John S. Baras and Shihwei Chen
Center for Satellite and Hybrid Communication Networks
Institute for Systems Research
A.V. Williams Building
University of Maryland
College Park, MD 20742
(301) 405-6606

Abstract

In this paper, we consider a large hybrid network which consists of a low-delay terrestrial sub-network and a high-bandwidth satellite sub-network. Both voice and data traffic are transmitted and routed through the same network. We show how to route both traffic via ground and/or satellite links. Two common voice/data integrated protocols such as fixed boundary and movable boundary schemes for the satellite channel are investigated, and the performance of both schemes is evaluated. The optimal splitting ratios for voice and data at the SIMPs (Satellite Interface Message Processors) are found using a powerful numerical optimization package (FSQP). We also develop dynamically optimal splitting ratios between satellite network and terrestrial network for data and voice traffic. We constantly monitor the traffic and measure the arrival rate and occupancy of every link. Based on these data, we optimize data traffic delay in a suitable time-frame under the constraints of voice traffic blocking probabilities of voice transmission links being less than specified value, which is up to system's design or users' requirements. A dynamic routing algorithm is presented.

INTRODUCTION

As networking technologies evolve at a steady and growing speed, the services available to users become more diversified. Multi-media information transmission is the goal of the next evolution of networks. Voice, data, images, graphics, etc., will be sent through networks at gigabits per second with high quality at users' demand. In addition to the traffic being heterogeneous, the transmission media can be hybrid. For example, the path of transmission can consist of a combination of ground and space routes. This kind of hybrid networks are the so-called mixed-media networks. Routing and flow control in such networks are important functions in achieving the goal of providing new services to users.

There have been several related works in the routing design of a mixed-media network (Abramson 1973 and Benelli 1983). Huynh et al. (1977) presented the optimal design of routing and capacity assignment in mixed-media packet-switched networks consisting of a ground subnet and a satellite subnet. In their algorithm, they assumed linear cost-capacity functions for both terrestrial and satellite links and a fixed-split routing policy in their terrestrial sub-network. The first assumption makes their capacity assignment problem mathematically tractable and thus a closed-form solution was obtained using analytic procedures involving Lagrange multipliers. The second assumption eliminates the routing problem within the terrestrial sub-network and reduces the flow assignment problem to one of determining the optimal amount of traffic which goes through satellite links for each pair of source/destination nodes.

More recently, Yuan and Baras focused on the control of mixed-media communication processors where the transmission of messages has a time constraint given by the user. They approximated the delay distribution by that of a product-form network model. Using a DFT algorithm, they computed a time index from the delay

CONF 950110

distribution so that 99 percent of messages arrived before the index. From this index, they derived the "Gittins index" which guides the processor to switch the packets to the subnetwork with the minimum discounted cost (delay).

However, these papers considered only one uniform transmission and switching mode, for example, only packet-switching for data transmission among source/ destination pairs. To consider a voice/data integrated system, we must modify the objective cost function to include the performance measure for voice traffic. In the telecommunication network, the blocking probability of the voice transmission is the major concern for phone-call establishments and should be minimized, because voice traffic must be transmitted in a continuous stream with very low variability of the time delay, while calls which do not get the resources to transmit are blocked and cleared. In contrast, data traffic, which may be either bursty or regular in nature, can be delayed and buffered for later transmission. It is the delay for packet transmission of data traffic that we want to minimize. Thus, the overall objective function of an integrated voice/data system is usually a weighted sum of blocking probability for voice traffic and delay for data traffic.

In our previous work (S. Chen 1992), we have obtained the optimal routing ratios for such mixed-media networks with integrated voice and data traffic using this type of objective function. However, the physical meaning of the weighting coefficients is not clear, except that a larger weight on delay means that data traffic is more of concern than voice traffic; smaller weight on delay means that data traffic is of less concern. There are no systematic procedures to find the weighting coefficients so far.

Due to these reasons, we believe that a different approach is necessary. Sometimes, the designer of the system may not want to optimize both data delay and voice blocking probability at the same time. The designer may just want to optimize the voice blocking probability or residual capacity without sacrificing the data traffic too much. Alternately the data traffic may have some real time constraints to meet. In these cases, it is more appropriate to optimize the performance objective for voice traffic under the constraints on data traffic link delays. Using similar arguments, we may alternately want to guarantee some link or trunk performance of voice traffic while optimizing the data traffic delay. In the former case, we must use a single objective (average network delay) and the constraints are that average voice blocking probabilities are less than some specified values. Since we can specify any link delay or trunk blocking probability, we have more detailed control over the performance of the network. In either case, we have a non-linear programming problem with non-linear constraints (S. Chen to appear).

MODEL

A mixed-media network could comprise several sub-networks. However, to simplify the problem, we consider a communication network composed of two sub-networks: one is the ground subnet, the other the satellite subnet. A mixed-media network with more than two sub-networks will be considered in the future, though the problem would become harder. The nodes are the locations of interface message processors (IMPs) linked together by landlines. There exist special nodes called satellite IMPs (SIMPs) which are interface message processors between the satellites and the ground links.

Routing in a mixed-media network consists of two major portions: (1) splitting of the input traffic at SIMPs between ground and satellite subnets; (2) routing on the ground subnet. The traffic of the ground subnet consists of voice traffic and data traffic. Our design problem can be stated as follows. Given a network topology, a ground routing procedure, and link capacities; we want to minimize the average system delay of data traffic under the constraints that the voice blocking probabilities of the ground links and the satellite channel be less than the specified performance requirements. We can consider other performance measures to be optimized depending on the application. For example, we may want to maximize the residual capacities of trunk groups with (or without) the delay constraints. Or we may minimize capital cost which is to be defined in different cases. What we present here is a general algorithm which can be used in different applications.

Let g_{ij} (s_{ij}) be the splitting factor of data traffic which specifies the fraction of data (voice) traffic, originating at node i and destined for node j, going through the ground sub-network. In the following, we will derive the overall objective function.

Data Delay in the Ground

Suppose that we are given a data sub-network in the ground of N nodes linked by L ground links of capacities C_{dl} (bits/sec), $l = 1, 2, ..., L$, in a specified topology. The network is partitioned into M regions, each having a SIMP. These M SIMPs are linked via a satellite channel of capacity C_s (bit/sec). A traffic rate matrix $[r_{ij}]$ specifies, in packets/sec, the average rates of messages flowing between all possible IMP pairs i and j, where $i, j = 1, 2, ..., N$.

Under the typical (Kleinrock independence) assumptions the average data delay in the ground (C. Wu 1988) is:

$$D_g = \frac{1}{\gamma} \sum_{l=1}^{L} \lambda_{dl} T_l, \quad T_l = \frac{1}{\mu_d C_{dl} - \lambda_{dl}} \tag{1}$$

where T_l is the delay on link l, $\gamma = \sum_{i,j=1}^{N} \gamma_{ij}$ = the total data traffic rate in the data sub-network, λ_{dl} = the traffic rate on link l, and $\frac{1}{\mu_d}$ = the average length of a packet.

Voice Blocking in the Ground

Suppose that we are given a telephone sub-network in the ground which may use the same transmission links and switching facilities of the data sub-network in the ground. This voice sub-network has N_1 nodes linked by L_1 trunks (links) of capacities $C_{vl}, l = 1, 2, ..., L_1$ in a specified topology. The capacity C_{vl} of link i can be divided into N_{vl} channels. A traffic rate matrix $[\Gamma_{ij}]$ specifies, in calls/min, the average rates of call requests between all possible IMP pairs i and j, where $i, j = 1, 2, ..., N_1$. We can model each trunk by an $M/M/N_{vl}/N_{vl}$ system and the average blocking probability P_b is (D. Bertseka 1987)

$$P_b = \frac{1}{\Gamma} \sum_{l=1}^{L_1} \lambda_{vl} P_l, \quad P_l = \frac{(\lambda_{vl}/\mu_v)^{N_{vl}}/N_{vl}!}{\sum_{n=0}^{N_{vl}} (\lambda_{vl}/\mu_v)^n/n!} \tag{2}$$

where P_l is the blocking probability of trunk l, $\Gamma = \sum_{i,j}^{N_1} \Gamma_{ij}$ = the total voice traffic in the voice sub-network, λ_{vl} = the traffic rate on link i, and $\frac{1}{\mu_v}$ = the average holding time of a phone call.

Satellite Channel

To integrate voice and data traffic in the satellite channel, we can have two strategies: a fixed boundary strategy or a movable boundary strategy. In the fixed boundary strategy, the data packets are not allowed to use the voice channels even if some of them are idle. In the movable boundary strategy, the data packets can occupy any of the voice channels not currently in use. However, an arriving call has higher priority to preempt the data packets serviced in the voice channels.

We make the following assumptions in our model:

1. The SIMPs collectively generate Poisson data traffic at rate Λ_d packets/sec and Poisson voice traffic Λ_v calls/min (excluding the retransmission due to collision). The overall transmission rate (the original rate plus retransmission rate) of data and voice traffic into the satellite channel are denoted as Λ'_d and Λ'_v respectively.

2. The data packets are of fixed length. The voice call duration is exponentially distributed with mean $1/\mu_v$ minutes.

3. Channels are slotted. Let T denote the slot length which equals the transmission time of a packet and S be the round-trip delay of the channel measured in slots.

4. The retransmission delay for a request or a random access data packet is uniformly distributed between 0 and K slots.

Data Delay in the Satellite Channel

There are R reservation channels and N_s message channels which are further divided as N_v voice channels and N_d data channels. The word "message" here refers to either voice calls or data packets. All channels are slotted. The length of a slot time is equal to the transmission time of a data packet. A time slot is further divided into n minislots for message reservation and the length of each minislot (T/n) is equal to the transmission of a request packet. There are N_s (nR) minislots in the R reservation channels. Among the N_s minislots, the first N_v are used for voice requests and the other $(N_s - N_d)$ are used for data requests.

Define p_{suc} to be the probability that a data packet will be successfully transmitted via a data channel. Then p_{suc} is $\frac{\Lambda'_d}{m}Tc^{-\frac{\Lambda'_d}{m}T}$. Here we assume that the data packets can go to any of the m data channels and the service at each channel is independent of other channels' services. The throughput of m (could be N_d in the fixed boundary scheme or a variable in the movable boundary scheme) data channels system η_{RAD} is $m \times p_{suc}$. The average delay under random access data (RAD) protocol is:

$$D_{RAD}(m) = [1.5 + S + (e^{\frac{\Lambda'_d}{m}T} - 1)t_{rx}]T \tag{3}$$

where $e^{\frac{\Lambda'_d}{m}T} - 1$ is the average number of retransmissions required for the data packet, $t_{rx} = 1.5 + S + \frac{K-1}{2}$ is the average retransmission time measured in slots, and $\Lambda_d = \Lambda'_d e^{-\Lambda'_s T}$.

For the demand assignment protocol, m is the number of data request minislots available for requesting a data channel in the system. Define p_{req} to be the probability that a data request packet will be successful on a minislot whose length is $\frac{T}{n}$. Then, $p_{req} = \frac{\Lambda'_d}{m}\frac{T}{n} \times e^{-\frac{\Lambda'_d}{m}\frac{T}{n}}$, and the throughput of the reservation channel η_{req} is $m \times p_{req}$. The delay of a successful request is $t_{req} = [1.5 + nS + (e^{\frac{\Lambda'_d}{m}\frac{T}{n}} - 1)t_{rx}]T/n$. The average delay for a data packet transmission is the sum of the request packet delay t_{req}, the queueing delay t_q and the propagation delay S.

$$D_{DAD} = [\frac{1.5}{n} + 2S + \frac{\Pi'(1)}{\Lambda'_d\frac{T}{n}e^{-\frac{\Lambda'_d}{m}\frac{T}{n}}} + (e^{\frac{\Lambda'_d}{m}\frac{T}{n}} - 1)t_{rx}]T \tag{4}$$

where $\Pi'(1)$ is the average queue length (J. Chang 1982), and $t_q = \Pi'(1)/\eta_{req}$, according to Little's formula, which is also negligible due to the same reason.

In this scheme, we note that the delay is at least "two hops". Therefore, it is not suitable for real time traffic.

Voice Blocking in the Satellite Channel

For voice channel, this is an $M/G/N_v/N_v$ system, and the average blocking probability P_s is given by

$$P_s = \frac{\Lambda_v}{\Gamma}P_v, \quad P_v = \frac{(\Lambda_v/t_v)^{N_v}/N_v!}{\sum_{k=0}^{N_v}(\Lambda_v/t_v)^k/k!} \tag{5}$$

where P_v is the blocking probability of the satellite channel, $\Lambda_v = \sum_{\sigma=1}^{M}\Lambda_{v\sigma}$ is the overall voice traffic rate into the satellite channel, $\Lambda_{v\sigma}$ =the voice call arrival rate from SIMP σ into the satellite channel, and $1/t_v$ is the average call duration plus the round-trip delay ST plus the call request and set-up time.

Fixed Boundary Scheme

Under the fixed boundary strategy, the data packets are not allowed to use the voice channel. The transmissions of voice calls and data packets do not affect each other. Thus, the performance analysis are the same as above.

Movable Boundary Integrated Protocol

The data packets can use the idle voice channels in this strategy. To simplify the calculations, we can assume that the data queues reach their stationary state when k, $0 \le k \le N_v$, voice calls are active. This is reasonable because the average call duration is much larger than call request and set-up time which includes propagation delay and random retransmission delay in the satellite channel. The average packet delay is

$$D_{MB} = \sum_{k=0}^{N_v} \pi_v(k) d_{data}(N_v - k) \tag{6}$$

where d_{data} is obtained from either Equation (3), or (4); N_v could be as large as N_s which is a constant. In this case, all the channels are used by the voice traffic and data traffic uses the channels not occupied.

Average Delay in the Satellite Channel

The overall average delay D_s in the satellite channel is

$$D_s = \frac{\Lambda_d}{\gamma} D_d \tag{7}$$

where D_d can be D_{RAD}, D_{DAD}, D_{MB}, or other analytical expressions of other data access schemes, $\Lambda d = \sum_{\sigma=1}^{M} \Lambda_{d\sigma}$, and $\Lambda_{d\sigma}$ = the data arrival rate from SIMP σ into the satellite channel.

OPTIMIZATION PROBLEM

We are given blocking probabilities of all links as constraints to minimize the average delay. The overall objective function which we want to minimize is thus given as

$$f(\lambda_{dl}, \lambda_{vl}, \Lambda_d, \Lambda_v) = D_g + D_s$$
$$\text{subject to} \quad 0 \le g_{ij} \le 1, \quad 0 \le s_{ij} \le 1, \quad P_l \le c_l, \ \forall l, \ P_v \le c_s \tag{8}$$

where P_l is the blocking probability of link l, c_l is the constraint of link l, P_v is the blocking probability of satellite link, and c_s is the bound of it. As we pointed out earlier, the objective function and constraints can be changed depends on the applications.

THE ALGORITHM

The algorithm is an adaptation of the algorithm described in (M. Schwartz 1987). This is applied to a typical mixed-media network like the one shown in Figure 1.

1. During the interval [t-T,t), the voice and data arrival rates of each link are measured by the local switches;
2. At time t, the measurements are sent to CCC;
3. At the CCC, we solve the optimization problem defined by Equation (9);
4. The results are sent to the appropriate local switches;
5. During [t,t+T), the switches send the incoming traffic to satellite channel accordingly.

Discussion

The measurements of the arrival rates in the previous interval are not the exact arrival rates for the present interval, but they are good estimates. We can apply Kalman filtering techniques to smooth out the estimates (M. Schwartz 1987).

There are two major considerations in the choice of T. The interval T must be small enough to capture and control the transients of the network. However, T can not be too small, otherwise it will cause instability of the network as we know from classical control theory. There are other considerations in picking the length of T. In our example, we can choose T to be on the order of half an hour.

CONCLUSION

We have proposed a dynamic optimal routing algorithm for hybrid networks with both voice and data traffic using an optimization based approach with multiple criteria. With a suitable update time-frame, we are able to adjust to the dynamics of the networks and find optimal switching ratios accordingly.

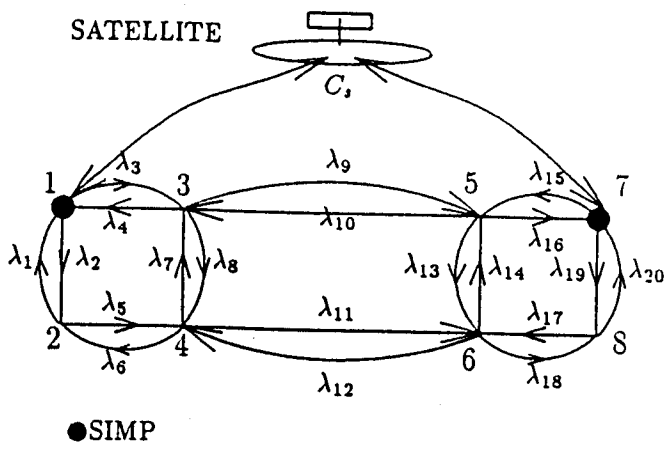


FIGURE 1. A Mixed-Media Network.

Acknowledgment

This work was supported in part by the Center for Satellite and Hybrid Communication Networks under NASA contract NASA NAGW-2777.

References

Abramson. N. (1973) "Packet Switching With Satellites," In *Proc. Nat. Computer Conf.*, 696-702.

Benelli G. and Del Re E. and Fantacci R. and Mandelli F. (1984) "Performance and Uplink Random-access and Downlink TDMA Techniques for Packet Satellite Networks", *Proc. IEEE*, 72 (11): 1583-1593.

Bertsekas D. and Gallager R. (1987). *Data Networks.* Prentice-Hall.

Chang J. and Lu L. (1982) "High Capacity Low Delay Packet Switching via Processing Satellite," In *Proc. Int. Commun. Conf. '82* vol. 1E.5.1-1E.5.5.

Chen S. and Baras J. S. (1992) "Optimal Routing in Mixed Media Networks with Integrated Voice and Data Traffic," In *Proc. GLOBECOM'92*, 335-339.

Chen S. and Baras J. S. "Tradeoff Curves in Mixed Media Networks with Integrated Voice and Data Traffic," *to appear.*

Huynh D. and Kobayashi H. and Kuo F.(1977) "Optimal Design of Mixed-Media Packet-Switching Networks: Routing and Capacity assignment," *IEEE Trans. Commun.*, vol COM-25(no. 1): 156-187.

Kheradpir S. (1990), "PARS: A Predictive Access-Control and Routing Strategy for Real-Time Control of Telecommunication Networks," *Proc. Network Management and Control, ed. by A. Kershenbaum*, 389-413.

Kleinrock L. (1964) *Communication Nets: Stochastic Message Flow Delay.* McGraw-Hill.

Schwartz M. (1987) *Telecommunication Networks: Protocols, Modeling and Analysis*, Addison Wesley.

Wu C. and Li V. (1988) "Integrated voice and data protocols for satellite channels". In *Proc. Mobile Satellite Conference, Jet Propulsion Lab*, 413-422.

Yuan J. (1988) *The Control of Multi-Media Communication Processors: For Messages with Time Constraints*, Dept. of Elect. Eng., U. of Maryland.